ENCYCLOPEDIA OF LIBRARY AND INFORMATION SCIENCE

Executive Editor

ALLEN KENT

SCHOOL OF LIBRARY AND INFORMATION SCIENCE
UNIVERSITY OF PITTSBURGH
PITTSBURGH, PENNSYLVANIA

Administrative Editor

CAROLYN M. HALL

ARLINGTON, TEXAS

VOLUME 61

SUPPLEMENT 24

MARCEL DEKKER, INC. NEW YORK • BASEL • HONG KONG

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

LIBRARY OF CONGRESS CATALOG CARD NUMBER 68-31232

ISBN 0-8247-2061-X

Current Printing (last digit):
10 9 8 7 6 5 4 3 2 1

PRINTED IN THE UNITED STATES OF AMERICA

43. F. C. Thorne, "The Citation Index: Another Case of Spurious Validity." *J. Clin. Psych.*, **33**, 1157–1161 (1977).
44. M. H. MacRoberts and B. R. MacRoberts, "Author Motivation for Not Citing Influences: A Methodological Note." *JASIS*, **39**, 432–433 (1988).
45. M. H. MacRoberts and B. R. MacRoberts, "Quantitative Measures of Communication in Science: A Study of the Formal Level." *Soc. Stud. Sci.*, **16**, 151–172 (1986).
46. C. G. Prabha, "Some Aspects of Citation Behavior: A Pilot Study in Business Administration." *JASIS*, **34**, 202–206 (1983).
47. S. M. Dhawan, S. K. Phull, and S. P. Jain, "Selection of Scientific Journals: A Model." *J. Doc.*, **36**, 24–41 (1980).
48. W. A. Satariano, "Journal Use in Sociology: Citation Analysis versus Readership Patterns." *Libr. Q.*, **48**, 293–300 (1978).
49. R. E. Stern, "Uncitedness in the Biomedical Literature." *JASIS*, **41**, 193–196 (1990).
50. E. Garfield, "Is Citation Analysis a Legitimate Evaluation Tool?" *Scientometrics*, **1**, 359–375 (1979).
51. C. D. Hurt, "A Comparison of a Bibiliometric Approach and an Historical Approach to the Identification of Important Literature." *Inform. Proc. Mgt.*, **19**, 151–157 (1983).
52. C. D. Hurt, "Important Literature in Endocrinology: Citation Analysis and Historical Methodology." *Libr. Res.*, **4**, 375–384 (1982).
53. P. Vinkler, "A Quasi-Quantitative Citation Model." *Scientometrics*, **12**, 47–72 (1987).
54. B. R. Boyce and C. S. Banning, "Data Accuracy in Citation Studies." *RQ*, **18**, 349–350 (1979).
55. J. H. Sweetland, "Errors in Bibliographic Citations: A Continuing Problem." *Libr. Q.*, **59**, 291–304 (1989).
56. G. A. Matter and H. Broms, "The Myth of Garfield and Citation Indexing." *Tidskrift for Dokumentation*, **39**, 1–8, 29 (1983).
57. S. Klimley, "Limitations of *Science Citation Index* in Evaluating Journals and Scientists in Geology," in *Proceedings of the 28th Meeting of the Geoscience Information Society*, Boston, Oct. 25-28, 1993, C. Wick, ed. Geoscience Information Society, Alexandria, VA, 1994, pp. 23–31.
58. B. Kefford and M. B. Line, "Core Collections of Journals for National Interlending Purposes." *Interlend. Rev.*, **10**, 35–43 (1982).
59. D. Pauly, "Who Cites Your Publications When You Work in the Tropics?" *ICLARM Newsl.*, **7**, 6–7 (1984).
60. J. MacLean, "Characteristics of Tropical Fisheries Literature." *ICLARM Newsl.*, **7**, 3–4 (1984).
61. R. E. de Bruin, R. R. Braam, and H. F. Moed, "Bibliometric Lines in the Sand" (commentary). *Nature*, **349**, 559–562 (1991).
62. M. Carpenter and F. Narin, "The Adequacy of Science Citation Index (SCI) as an Indicator of International Scientific Activity." *JASIS*, **32**, 430–439 (1981).
63. R. B. Archibald and D. H. Finifter, "Bias in Citation Based Ranking of Journals." *Schol. Pub.*, **18**, 131–138 (1987).
64. P. Moorbath, "A Study of Journals Needed to Support the Project 2000 Nursing Course with an Evaluation of Citation Counting as a Method of Journal Selection." *Aslib Proceed.*, **45**, 39–46 (1993).
65. R. Taylor, "Is the Impact Factor a Meaningful Index for the Ranking of Scientific Research Journals?" *Can. Field Naturalist*, **95**, 236–240 (1981).
66. J. L. Kelland, "Biochemistry and Environmental Biology: A Comparative Citation Analysis." *Libr. Inform. Sci. Res.*, **12**, 103–115 (1990).
67. C. Tomer, "A Statistical Assessment of Two Measures of Citation: The Impact Factor and the Immediacy Index." *Inform. Proc. Mgt.*, **22**, 251–258 (1986).
68. R. E. Rice, C. L. Borgman, D. Bednarski, and P. J. Hart, "Journal-to-Journal Citation Data: Issues of Validity and Reliability." *Scientometrics*, **15**, 257–282 (1989).
69. M. H. MacRoberts and B. R. MacRoberts, "Problems if Citation Analysis: A Critical Review." *JASIS*, **40**, 342–349 (1989).
70. B. C. Peritz, "On the Objectives of Citation Analysis: Problems of Theory and Method." *JASIS*, **43**, 448–451 (1992).
71. M. H. MacRoberts and B. R. MacRoberts, "Testing the Ortega Hypothesis: Facts and Artifacts." *Scientometrics*, **12**, 293–295 (1987).

72. J. R. Cole and S. Cole, "The Ortega Hypothesis." *Science*, **178**, 368 (1972).
73. W. C. Snizek, "In Search of Influence: The Testing of the Ortega Hypothesis." *Scientometrics*, **12**, 311–314 (1987).
74. S. Cole and J. R. Cole, "Testing the Ortega Hypothesis: Milestone or Millstone?" *Scientometrics*, **12**, 345–353 (1987).
75. H. Small, "The Significance of Bibiliographic References." *Scientometrics*, **12**, 339–341 (1987).
76. V. V. Nalimov, "Scientists Are Not Acrobats." *Scientometrics*, **12**, 303–304 (1987).
77. D. Lindsey, "Using Citation Counts as a Measure of Quality in Science: Measuring What's Measurable, Rather Than What's Valid." *Scientometrics*, **15**, (3–4) 189–203 (March 1989).
78. A. J. Nederhof and A. J. Van Raan, "Citation Theory and the Ortega Hypothesis." *Scientometrics*, **12**, 325–328 (1987).
79. S. M. Lawani, "The Ortega Hypothesis, Individual Differences and Cumulative Advantage." *Scientometrics*, **12**, 321–323 (1987).
80. F. E. DeHart and L. Scott, "ISI Research Fronts and Online Subject Access." *JASIS*, **42**, 386–388 (1991).

JOHN LAURENCE KELLAND
ARTHUR P. YOUNG

COMPUTER SUPPORTED INDEXING: A HISTORY AND EVALUATION OF NASA'S MAI SYSTEM

Introduction

Computer supported indexing systems may be categorized in several ways. One classification scheme refers to them as statistical, syntactic, semantic or knowledge-based. While a system may emphasize one of these aspects, most systems actually combine two or more of these mechanisms to maximize system efficiency (*1, 2*).

Statistical systems can be based on counts of words or word stems, statistical association, and correlation techniques that assign weights to word locations or provide lexical disambiguation, calculations regarding the likelihood of word co-occurrences (*3*), clustering of word stems and transformations, or any other computational method used to identify pertinent terms. If words are counted, the ones of median frequency become candidate index terms.

Syntactical systems stress grammar and identify parts of speech. Concepts found in designated grammatical combinations, such as noun phrases, generate the suggested terms.

Semantic systems are concerned with the context sensitivity of words in text. The primary goal of this type of indexing is to identify without regard to syntax the subject matter and the context-bearing words in the text being indexed (*4*).

Knowledge-based systems provide a conceptual network that goes past thesaurus or equivalent relationships to knowing (e.g., in the National Library of Medicine (NLM) system) that because the tibia is part of the leg, a document relating to injuries to the tibia should be indexed to LEG INJURIES, not the broader MeSH term INJURIES, or knowing that the term FEMALE should automatically be added when the term PREGNANCY is assigned, and also that the indexer should be prompted to add either HUMAN or ANIMAL (*5*).

Another way of categorizing indexing systems is to identify them as producing either *assigned-* or *derived-*term indexes.

> An *assigned-term* index is provided by an indexer who uses some intellectual effort to determine the subject matter of the document at hand, and assigns descriptors from a controlled vocabulary to identify the concepts expressed by the document's author.
> A *derived-term* index uses descriptors taken from the item itself (*6*). One kind of a derived-term index is an index found in the back of a book.

The National Aeronautics and Space Administration's (NASA's) Center for AeroSpace Information (CASI) indexes technical reports using a machine-*aided* indexing (MAI) system that was originally syntactic. Today it is primarily semantic and computational. It has been designed as a computer *aid* for indexers. Emphasis is placed on the word *aided* in NASA's MAI system because all output is expected to be reviewed. The NASA/CASI indexers do some back-of-the-book, *derived*-term indexing for a few special documents, but they primarily index technical reports with *assigned* NASA thesaurus terms, many of which are suggested by MAI.

The NASA MAI System

NASA's MAI system is fully operational and cost-effective. It started with a third generation of the Defense Technical Information Center's (DTIC's) original syntactic system, and by 1996 was using a third generation of NASA's first system. MAI was developed at NASA as part of a concentrated effort to speed up the indexing of scientific and technical reports and cut costs. MAI functions within normal NASA time constraints and workloads, and is used in conjunction with an electronic input processing system (IPS).

The NASA MAI system was changed from syntactic to semantic in order to make processing fast enough for an on-demand, online, interactive system—which is available now in addition to the standard batch processing. However, processing speed was not the only reason for choosing a semantically based design over a syntactic one. There are several other arguments, such as (1) the large number of rules required for a syntactic-based system to handle different meanings of context-sensitive words, (2) the enormous amount of information needed to disambiguate words, and (3) the attention of syntactic systems to form rather than content (*7*). NASA's present system is based on the co-occurrences in parts of a sentence of domain-specific terminology; that is, words and phrases that are not broad in their meanings, but that have (or suggest) domain-specific, semantically unambiguous, indexable concepts (*8*).

While the NASA/CASI system is largely semantic, according to the definition above, it also has computational aspects. Statistics are used to determine the probability of an indexer using a particular term when a given word or phrase is encountered in text. Statistics are used to determine which authorized posting terms will be targeted for identifying new knowledge base (KB) entries. Also, statistics were used in making the decision to limit the number of words between two concatenated words to a

maximum of three words. The current method of selecting KB entries is based on a statistical analysis of the single- and multiword phrases that occur in large volumes of text (*9*). These phrases occur in text that (1) resides in the NASA database, (2) is indexed to a targeted thesaurus term, and (3) contains the candidate words or "phrases" with relative frequency.

In addition to these computational aspects of its MAI system, NASA/CASI now calls its lexical dictionary or translation table a KB because of its conceptual network properties. While NASA's KB is not as sophisticated as NLM's, it still provides more information than just equivalent thesaurus terms. The NASA KB has entries that represent decisions regarding the relevancy of particular concepts (*9*). For example, within the aeronautics domain, the concept AIRCRAFT is much too broad in meaning to be a useful indexing term for most instances in which the word *aircraft* appears in text. In this case, specific entries in the KB would initiate a search for a multiword semantic unit such as A-320 AIRCRAFT, which describes the specific vehicle in question; or AIRCRAFT STABILITY, AIRCRAFT CONSTRUCTION MATERIALS, or AIRCRAFT CONFIGURATIONS, which indicate the particular aeronautical aspect of interest. Other entries in the KB serve to disambiguate certain words (such as *matrices*) which might refer to either mathematical matrices or material matrices. The KB disambiguates meanings with its choice of entries for the KB. Phrases or word strings, of course, may be selected now from semantically rich verbs and other parts of speech that do not occur in noun phrases. The process of identifying KB entries is similar to the one described by N. Vleduts-Stokolov for specifying "concept codes" from word co-occurrences in the BIOSIS database (*10*).

History

DTIC'S ROLE IN NASA'S MAI SYSTEM

Paul Klingbiel, first director of NASA's MAI project, was active for eighteen years in linguistic research at DTIC, formerly called the Defense Documentation Center (DDC). While there, he initiated a lexical dictionary that became part of DTIC's MAI system. Contrary to F.W. Lancaster's remark in his book *Indexing and Abstracting in Theory and Practice* (*11*), DTIC's lexical dictionary MAI system suggests to the indexers the same kinds of descriptors from the DTIC controlled vocabulary that human indexers assign. Indexers either approve or reject these terms and may add additional terms.

DTIC's first MAI system was established in the late 1970s. It was a phrase delineation method that sought to identify noun phrases for translation into controlled vocabulary terms. This system used a recognition dictionary, which assigned syntax to each word encountered in text; a machine phrase selection (MAPS) program, which strung words together according to specified grammar rules; and a kind of use reference file called the natural language data base (NLDB), which had as its core vocabulary the DDC thesaurus terms, excluding related and hierarchical terms (*12*). This system required that the entire phrase identified by MAPS be located as a key to an entry in the NLDB. Natural language phrases with a maximum length of four

words were added from MAI production runs when they did not match an entry already in the NLDB.

Between 1974 and 1979, about 250,000 natural language phrases were added to the core terms already in the NLDB, and the file became very large and cumbersome. The available manpower was not sufficient to cope with the large number of phrases produced by MAI. Projections indicated that the NLDB would at least double in size before the number of new candidate phrases substantially decreased. When it was determined that a final total of a million phrases was quite possible, building an NLDB was abandoned in favor of a new, more compact structure call the lexical dictionary (*13*).

After retiring from DTIC, Klingbiel was persuaded to work for a year at the NASA Center for AeroSpace Information (CASI—then the NASA Scientific and Technical Information Facility) to organize an MAI system for NASA. He brought with him the DTIC system's concept of using a lexical dictionary with a new feature added—a logic code. Logic codes resulted from an "aha" experience that occurred to Klingbiel after retirement. Logic codes provided information that reduced the number of KB entries that the computer had to review in order to find an entry that matched input. This reduced the number of records that the computer was obliged to read, and therefore reduced MAI processing time. Copies of DTIC's programs and prints of its lexical dictionary were obtained and studied, but could not be used directly because computer languages and equipment at the two agencies were not compatible. DTIC's programs were written for a UNIVAC mainframe and sent to NASA in COBOL, while NASA's programs were written in PL1 for an IBM mainframe.

By inverting DTIC's lexical dictionary information, a tape was obtained that showed how the NASA lexical dictionary system's KB could translate DTIC's thesaurus terms. The inverted tape was helpful as well in identifying natural language phrases that could be translated into NASA posting terms.

NASA KWOC AND DATA ENTRY

The DDC lexical dictionary was built from MAI production output. NASA's KB has been constructed from a variety of sources. Klingbiel began building the KB with a list of NASA thesaurus terms in a special key words out of context (KWOC) format. A KWOC listing had been used at DTIC to review and correct inconsistencies that had entered into its natural language database. By starting the KB with a KWOC printout of all of NASA's posting terms and use references, the problems experienced at DTIC were avoided. However, it was determined later that an alphabetical list of NASA terms would have worked just as well. The use of the KWOC was described in detail in NASA Contractor Report 3838 (*14*).

Each authorized posting term and use reference that appeared in the *NASA Thesaurus* was given an appropriate logic code, coded for keypunching and data entry, any entered into the KB. Completion of this first KB building phase had two results: (1) it established the capability for automatically translating or subject switching (SS), any DTIC posting term that *exactly matched* an authorized NASA term, and (2) it precipitated a decision to separate SS files and procedures from those files and programs that translate *natural language* words and phrases to authorized NASA

terms. Exactly matched meant that not only was there a character-by-character match of the DTIC and NASA terms, but also the meanings and uses were identical. For example, the term PERFORMANCE TESTS appears in both agencies' thesauri, but they do not exactly match because NASA uses this term only for machinery, whereas DTIC uses the term for animals or people as well as for machinery. The SS of all DTIC terms to NASA terms became operational in June 1983 and was fully described in NASA Contracter Report 3838 (*14*). During the following year a similar SS project was undertaken for translating to equivalent NASA thesaurus terms the authorized posting terms of the Department of Energy (DOE). This was a much larger task, and while never totally completed, the SS system was able to translate virtually all of the DOE terms that NASA encountered. The omissions were largely highly specific atomic energy terms and entries for coordinated DOE terms. In 1995 a SS table was constructed for yet another controlled vocabulary. In the meantime, the DTIC and DOE SS files were abandoned in favor of regular MAI. This was done not only to reduce file maintenance but also to improve indexing quality when one agency suspended indexer review. NASA MAI yielded better results than machine translation of other agency's machine indexing.

In order to do MAI of *natural language text*—as opposed to SS of another agency's controlled vocabulary—NASA first used a version of DTIC's system of identifying indexable concepts by parsing and selecting only noun phrases from text. This method is described in NASA Contractor Report 4512, *Machine Aided Indexing from Natural Language Text* under "SYSTEM DESCRIPTION: Text Processing with Access-1" (*12*). MAI of natural language text became operational initially for a single file in August 1986, and was made available as an online, interactive system for documents without abstracts in October 1988. At that time, it was determined that documents with abstracts took too long for online use of this system, requiring a wait of an average of 90 seconds for MAI-suggested terms.

PARSING ELIMINATED

A new method of identifying indexable concepts was needed to eliminate parsing, to use information not contained in noun phrases, and especially to shorten response time. An effort was made to use computation to identify semantic units, and a new program, called Access-2, was devised. Semantic units were identified by ordered concatenations of words within an arbitrarily established proximity leading to appropriate entries in the KB. The semantic unit in the NASA system is normally limited to a maximum of five words to ensure grammatically correct word associations without parsing; however the system can handle longer units if the words are consecutive. Search keys of fewer than five words must be created from within a five-word segment of the machine-selected string. This five-word proximity limit was established empirically and represents the best trade-off between identifying the most semantic units while limiting the risk of inappropriate word concatenations. The new semantic computational method became operational in May 1989 in an overnight batch mode for NASA's analytic Scientific and Technical Aerospace Reports (STAR), in a daytime batch mode for other STAR documents in March 1990, and for all of the main document series in an online, interactive mode in June 1990. The response time was reduced from 90 seconds to about 6 seconds with the elimination of parsing.

A computational method for identifying new KB entries also was devised from analyses of text targeted to specific thesaurus terms. This knowledge base building (KBB) activity uses a text analysis tool that operates—usually—on the titles and abstracts of a large set of records (150–1200) that are indexed to, or otherwise identified as being related to, a single thesaurus concept. The text is processed to identify all possible one-, two-, three-, four-, and five-word "phrases" that might be created by MAI programs. These phrases or word combinations are filtered syntactically to prevent prepositions and articles from occurring at the beginning or end. The phrases are sorted by the number of words that they contain, and within that sort, by the frequency of occurrence in the body of text that is being examined. Frequently occurring phrases that are synonymous to the original index term are added to the KB (*15*). Whatever procedure is used to identify new KB entries, the intent is to build a KB sufficiently comprehensive to translate whatever natural language text is input to an equivalent output in NASA thesaurus terms in such a manner that the indexers' role is largely editorial.

CLIENT SERVER ENVIRONMENT

In August 1991, NASA and its CASI contractor, RMS Associates, made the decision to transfer its operations from a mainframe to a client server environment. Client server architecture was determined to be better than mainframe equipment for sharing CASI resources. A database management system and the client server architecture were selected and in 1993 RMS began designing applications.

Improved capabilities of client server architecture have been made possible by advancements in the power, speed, and miniaturization of chips, networking technologies, personal computer (PC) storage capacity, and interface development tools. Client server systems also generally have lower capital investment and maintenance costs than mainframe systems (*16*). CASI's new client server system, which first became operational in October 1995, relies on an ORACLE database for its data repository. It allows a single source of information for multiple users, reduces the amount of data redundancy, and results in data that are more reliable and accurate.

This move away from mainframe support has required new programs in different languages and they are still at this writing being "polished." MAI applications and the KB access programs interface with a new IPS. A new KB editing system for the client server architecture was written originally in C++, but was replaced soon after with a less cumbersome system that uses a Delphi package and is written in Pascal. MAI response time for processing an average abstract through MAI in an interactive mode is from 3 to 9 seconds. This is on a 486 type 33 MHz PC with 16 million bytes of RAM running on Windows 3.1. The Oracle server is accessed through a Novell network.

MAI System Components

NASA's online MAI system has three components. The first is an *application program* that indicates the input text to be processed; selects text strings from the specified text; "calls" Access-2 and feeds those strings to it; accepts and stores NASA terms from Access-2; prints out various reports; and for NASA's electronic IPS,

provides an online display of NASA terms. The application programs differ for each use of MAI. The reports printed usually consist of a list of natural language words and phrases selected from the strings by Access-2 with their equivalent NASA thesaurus terms, and a list of words not found in the KB (the "third component," described below).

The second component is *Access-2,* a modular program that never acts by itself, but always is "called" by an applications program. Access-2 accepts strings from the applications program and puts each word in the string into its own array cell; examines these words from left to right in five-word segments beginning with word 1 and word 2; and constructs potential keys or semantic units that are used for searching the KB. (See Fig. 2.) The program then compares the first word in each of these search keys with the first word in the keys in the KB to see if that word exists. If it does not exist, the word is stored in a list of "words not found as first word in a key," sent back to the application program, displayed on the IPS indexing screen for indexer review, and, in batch processing, printed out for consideration as a new KB entry. These unfound words are also displayed for the quality assurance staff. If the first word and the second word are found as a key to a record in the KB, then the posting term field is read. Any record that contains one or more posting terms will have that or those thesaurus terms as output to the user. If the potential key is found and an asterisk is in the posting term field, the program will look for another word that added to the first two words will produce a key that leads to postings. A key that is found and has zeros in the posting term field will not be translated, but will be flagged so that the words in the key cannot be used again in another key without adding a previously unused word to it or them.

The third component is the *knowledge base*. The KB contains the vocabulary, relationships, and rules surrounding the vocabulary (*17*). In the NASA CASI mainframe system, this is stored in a dataset that provides thesaurus term equivalents for input natural language words or word combinations. It also normalizes concepts that are expressed in different ways. It should be noted that the KB fields (i.e., the key field and posting term field) comprise a more robust rewrite system than that of use references in a standard thesaurus. The usual use reference types (with the addition of semicolons and a 999 flag, as explained below) are included as keys to authorized terms, but a new and very powerful concept is added—the rewrite to 00 in the posting term field. Linguistically, this deletion is a zeroing rule. It suppresses unwanted translations of natural language. Additionally, the KB directs the computer to look for word combinations of more than two words when they exist in the KB. At this writing the KB is still available on the mainframe where it is stored in a virtual storage access method (VSAM) file that contains more than 121,000 records—or rules.

Two fields are essential to NASA's KB. One is the *key field* of each record, which is unique and serves as the computer address to the entry in the KB, and the other is the *posting term field.*

The key field consists of one of the following three combinations:

1. Any word followed by a semicolon and three nines. (Nines are used because they sort last in NASA's IBM-4381 mainframe on which MAI has been processed.) A single word followed by at least two of whatever character sorts last is wanted because that entry is the default lookup and is of interest only when other combinations beginning with the initial word are not found. The word combinations beginning with the same initial word are searched sequentially in the computer's sort order. Sort order on the IBM mainframe

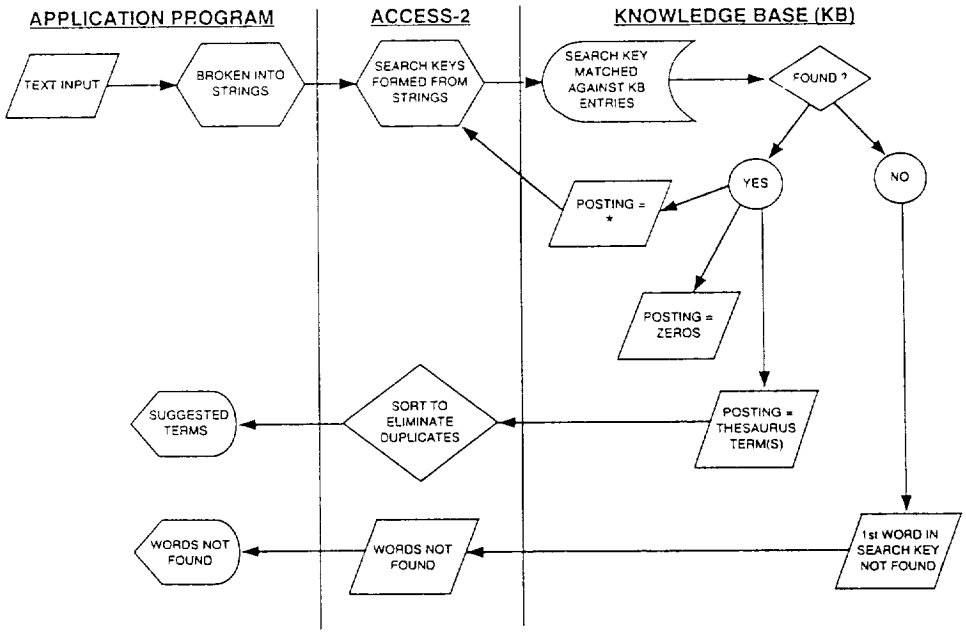

FIGURE 1. *NASA's online machine-aided indexing.*

begins with space and symbols, follows with alphas, and ends with numbers 0 through 9. In the original DTIC system, because zeros sorted after the nines and therefore were last on DTIC's computers, the key that contained only one word ended with ";00."

2. A second possible combination for a key is two or more words separated by semicolons.
3. A third key combination consists of two or more words separated by semicolons followed by another semicolon and three nines (";999"). This combination is required when a two-word combination is imbedded in a longer key and a translation is wanted for those two words when the longer key cannot be found.

The posting term field also has three possibilities for its contents. It may hold: (1) one or more thesaurus terms that are equivalent in meaning to the key for that entry; (2) two zeros (00), which indicate that no translation for that key is wanted; or (3) an asterisk (*), which indicates that the computer should look for an additional word that will make a longer and more specific key.

Functions of the MAI Components

The functions of the MAI system's components are illustrated in Fig. 1, entitled "NASA's Online Machine-Aided Indexing." The process shown in Fig. 2 illustrates what Access-2 does to form search keys from word strings. As indicated in the

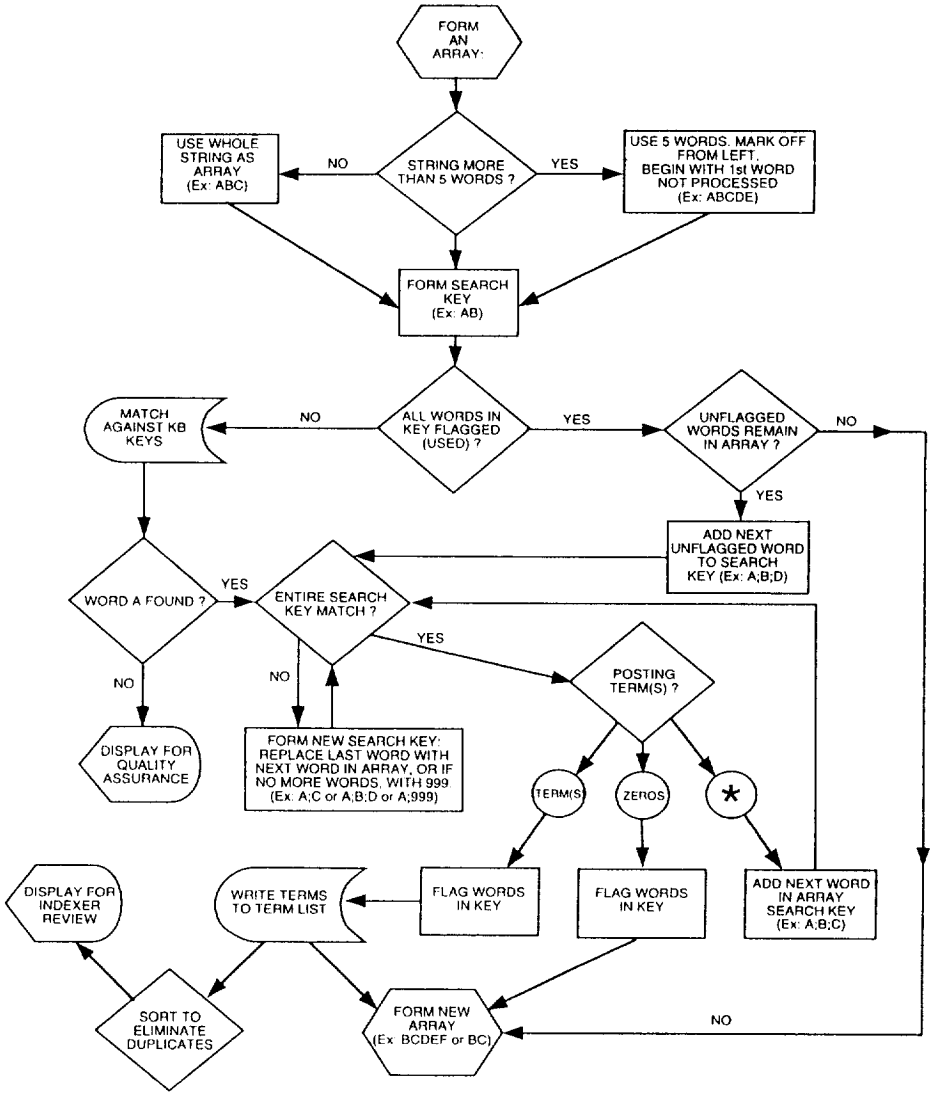

FIGURE 2. *To form search keys from word strings.*

hierarchy shown in Fig. 3, the Access-2 program breaks text into word strings. Five-word arrays are identified from which potential keys to thesaurus terms are constructed for searching the KB.

For an array of "words" A B C D E, the possible search key combinations are as listed in Fig. 3, which shows search key construction.

For example, if word A followed by word B is not found in the KB, then word A

Text

. Word Strings

. . 5-Word Arrays

. . . Search Keys

For an array of "words" A B C D E, the possible search keys are:

A;B	A;B;C	A;C;D
A;C	A;B;C;D	A;C;D;E
A;D	A;B;C;D;E	A;C;D;999
A;E	A;B;C;D;E;F	A;C;E
A;999	A;B;C;D;E;F;G	A;C;999
	A;B;C;D;E;F;999	
	A;B;C;D;E;999	
	A;B;C;D;999	
	A;B;C;999	A;D;E
	A;B;D	A;D;999
	A;B;D;E	
	A;B;D;999	
	A;B;E	
	A;B;999	A;E;999

FIGURE 3. *Search key construction.*

followed by word C is looked up. If that is not found, the search key becomes word A followed by word D, and if that is not found, the program looks for word A followed by word E. If none of these are found, word A (A;999) is looked up by itself. These possible search keys are listed sequentially in column 1 in Fig. 3.

On the other hand, if word A followed by word B is found and an asterisk is in the posting term field, the possible search key combinations are listed in column 2 of Fig. 3. Note that if five consecutive words A;B;C;D;E are found with an asterisk in the posting term field, the program will look for the next word in the string. Likewise, if six consecutive words A;B;C;D;E;F are found with an asterisk in the posting term field, the program will add the seventh or G word from the string. The longest existing key contains seven words, but longer keys can be used if deemed necessary.

If a word A followed by word C or D or E has an asterisk in the posting term field, the possible search keys are listed in column 3 of Fig. 3.

When one designs an MAI system, the procedures selected for its initial phrase delineation and analysis define what kinds of information needs to be represented in KB entries, and also how large an operational file will need to be. For example, the use of word stemming or phrase normalization could reduce the number of required entries. Likewise, the strategies used for disambiguating words and for analyzing relevancy can define the level of complexity required for knowledge representation and ultimately may dictate the kind of data structure that is used. In the particular case

of the NASA KB, when the trade-offs were considered, it was decided to keep all rules as simple as possible in order to keep the system's online response time as short as possible. By rules, we mean "if...then" statements. For examples: if "In-102" is encountered in the title or abstract, then provide the thesaurus term "INDIUM ISOTOPES" as a suggested term for indexer review; or if a word is hyphenated, then look in the KB for the hyphenated form. If it is found, then read the posting term field; or else (if it is not found) drop the hyphen and treat the hyphenated word as two separate words. Most rules in the NASA MAI system specify: (1) if the search key is found and the posting term field contains NASA thesaurus terms, then suggest the NASA thesaurus term(s) for review by the indexer; (2) if the search key is found and the posting term field contains an asterisk, then add the next word in the five-word array to the search key and look up the new search key; or (3) if the search key is found and the posting term field contains two zeros, then no translation to NASA thesaurus terms is wanted for that word or word combination.

Some MAI systems have more numerous rules than the NASA system. For example, they will examine instances of capitalization of words in the key or look for specific words in close proximity to a word in the key as part of the if statement (*18*). For example, if the word titanic occurs, and if it begins with an uppercase T, and if the word ship occurs within four words of Titanic, then return the term U.S.S. Titanic for indexer review. NASA system designers chose to eliminate as many details as possible in order to minimize the computer's read and write requirements and thereby maximize processing speed.

Another kind of computer-supported indexing system does a statistical count of keywords found in bibliographic references. The theory is that the salient concepts will be prominent in the titles of these references. For any scientific project that is on the cutting edge of research and development, this may not be the best system to choose, for it takes a sizeable body of material to make such a system work satisfactorily. Cutting-edge science is more likely to have references that treat the subject in question only peripherally. After all, if no one has done "it" before, how can they write about it?

Evaluation Measures

Machine-aided indexing was developed at NASA CASI in a high-pressure, production environment. Measurements of its results were devised not to be disruptive of the regular work flow. The following observations were made:

- In 10 years, the indexing staff decreased from eight to five people.
- The workload per person approximately doubled during the same period.
- Indexing is more consistent between indexers than it was before MAI. (This was noted by the person who has trained most of the present staff.)
- Fewer errors of omission are made. (Also noted by the trainer.)
- Less research time is required because of the expert advice provided by MAI as to appropriate technical terms.

It is reasonable to conclude from the above that the indexers, supported by the

NASA MAI system, have been able to maintain and even improve indexing quality, and at the same time increase production. However, MAI is not the only change responsible for increased productivity. Input is now done at a computer terminal or with a scanner or electronically from magnetic tape instead of with pen and paper. This also speeds processing. Other variables that can affect the measures of the system include the following:

- The amount of time available to an indexer. MAI terms may be questioned less if the workload is heavy and more if the load is light.
- The existence of similar terms; for example, SIMULATORS and SIMULATION. The indexer may select the term for the equipment described, while MAI may suggest the process.
- Valid terms that were suggested by MAI, appropriate for the document at hand, but not assigned.

It was determined in an early test with experienced indexers that MAI saved an average of three minutes per document by reducing the time needed to look up terms in the thesaurus (*14*). It is reasonable to expect that this time savings is even greater for comparatively new indexers who have not become thoroughly familiar with the variety of terms in NASA's controlled vocabulary.

MATCH RATE

Another early measure of how well the MAI system performed was referred to as the match rate. This term originally was used to describe the percentage of machine-selected words or phrases (semantic units) that could be found either entirely or partly in the key field of the KB. When that percentage reached the upper 90s, it lost its value as a measure of progress, and so it was redefined.

The match rate now refers to the percentage of MAI-suggested terms that the indexer elects to use. This measure, which began at 23 percent in early 1996 ranged from 40 percent to 60 percent—or an average of 50 percent—and it has risen gradually over the lifetime of the system as improvements have been made to it.

CAPTURE RATE

In 1986, NASA instituted a measure referred to as the capture rate. This describes the percentage of indexer-assigned terms that are suggested by MAI. The capture rate has been, rather consistently, a few percentage points higher than the match rate. Some systems refer to this measure as "hits."

CONSISTENCY FACTOR

In late 1989, we began to calculate the consistency (or quality) factor q. This identifies the percentage of common terms c found in two lists of terms, one generated automatically and represented by a, and the other terms selected intellectually by the indexer and represented by I. Expressed in another way, q is the ratio of the common terms to the unique terms, where $q = c/(a+i)-c$ (Refs. *19, 11*).

The following table shows the match rate, capture rate, and consistency factors calculated for 1987, 1988, and estimated for 1993.

Year	Match rate (%)	Capture rate (%)	Consistency factor (%)
1987	32.4	36.9	20.8
1988	37.0	39.0	23.4
1993	50.0	50.0	33.3

The 1987 figures are from a sample of approximately 2,500 documents. The 1988 figures were based on a sample of 100 documents, and the 1993 figures are from a survey of the indexers that were using the system.

As stated above, tests used to evaluate NASA's MAI system were limited to those that would not slow production. A very early test determined that approximately three minutes per document were saved by using MAI, and generally several more index terms were assigned when indexing was done with computer help. Perhaps the best proof of the success of MAI is that indexers handle more work than ever before, they like MAI, and there has been no adverse effect on retrieval evidenced by user or retrieval analysts' feedback.

Conclusions

The jury is still out on the case of the most efficient way to support information retrieval with, or even without, indexing. Parsing frameworks have become cleaner and more flexible (*2*). More studies in computational linguistics are being undertaken. Automatic parsers and KBs are becoming more numerous and more sophisticated. The application of standard generalized markup language (SGML) to electronic documents available on the Internet is facilitating the exchange of information. Some organizations index documents with terms derived from full text. Some indexing has human review and some is entirely automated. During the 1980s the trend in information processing was toward making retrieval systems more user friendly. In the 1990s the primary concern seems to be making the systems cheaper. Cheaper often means less user friendly systems and shifting some of the work of information discovery onto the searchers.

For an online system designer, quick responses have high priority. Regardless of the specific design selected for a MAI system, its overall performance is largely dependent upon the quality and the comprehensiveness of its KB. Strict control and input from domain experts are critical during the database development process. The time and other resources spent in careful construction of the KB pay off with high-quality output and indexer acceptance.

Acknowledgments

This work has been supported by NASA under contract NASw-4584. Appreciation also is extended to senior indexer Michael T. Genuardi and director of operations and analysis Gail M. Hodge at the NASA Center for AeroSpace Information, and Paul H. Klingbiel, consultant, for their comments and suggestions.

REFERENCES

1. J. L. Milstead, "Natural Language Processing," in *Automated Support to Indexing,* National Federation of Abstracting and Information Services seminar, Philadelphia, Oct. 14, 1993.
2. K. Spark Jones and Y. Wilks, *Automatic Natural Language Parsing,* Halstead Press, New York, 1983.
3. J. A. Guthrie, L. Guthrie, Y. Wilks, and H. Aidinejad, "Subject-dependent Co-occurrence and Word Sense Disambiguation," in *The 29th Annual Meeting of the Association for Computational Linguistics: Proceedings of the Conference,* University of California, Berkeley, June 18–21, 1991, pp. 146-152.
4. H. Borko and C. L. Bernier, *Indexing Concepts and Methods,* Academic Press, New York, 1978, pp. 113–137.
5. S. M. Humphrey, "Knowledge-based Systems for Indexing" in *Challenges in Indexing Electronic Text and Images,* R. Fidel, T. Bellardo Hahn, E. M. Rasmussen, and P. J. Smith, eds. published for American Society for Information Science by Learned Information, Inc., Medford, NJ, 1994, pp. 161–175.
6. D. B. Cleveland and A. D. Cleveland, *Introduction to Indexing and Abstracting,* 2nd ed., Libraries Unlimited, Englewood, CO, 1990.
7. B. J. Dorr, *A Lexical Conceptual Approach to Generation for Machine Translation,* Office of Naval Research, Arlington, VA, NTIS no. AD-A197356, Jan. 1988.
8. A. K. Melby, "Benefits and Limitations of Formal Systems in Technical Writing," in *Proceedings Second International Congress on Terminology and Knowledge Engineering, TKE'90: Terminology and Knowledge Engineering, Volume 1,* H. Czap and W. Nedobity, eds. Indeks Verlag, Frankfurt/M., Federal Republic of Germany, Oct. 1990, pp. 24–25.
9. M. T. Genuardi, "Knowledge-based Machine Indexing From Natural Language Text: Knowledge Base Design, Development, and Maintenance," in *Proceedings Second International Congress on Terminology and Knowledge Engineering, TKE'90: Terminology and Knowledge Engineering, Volume 1,* H. Czap and W. Nedobity, eds. Indeks Verlag, Frankfurt/M., Federal Republic of Germany, Oct. 1990, pp. 345–351.
10. N. Vleduts-Stokolov, "An Automatic Support to Indexing a Life Sciences Data Base." *Inform. Proc. Mgt.,* **18**(6), 313–321 (1982).
11. F. W. Lancaster, *Indexing and Abstracting in Theory and Practice,* University of Illinois, Champaign, 1991, pp. 60–85.
12. J. P. Silvester, M. T. Genuardi, and P. H. Klingbiel, *Machine Aided Indexing from Natural Language Text,* NASA Contractor Report 4512, National Aeronautics and Space Administration, Washington, DC, NTIS no. N93-26901, Oct. 1993.
13. P. H. Klingbiel, "Phrase Structure Rewrite Systems in Information Retrieval." *Inform. Proc. Mgt.,* **21**(2), 113–126 (1985).
14. J. P. Silvester, R. Newton, and P. H. Klingbiel, *An Operational System for Subject Switching Between Controlled Vocabularies: A Computational Linguistics Approach,* NASA contractor report no. 3838, National Aeronautics and Space Administration, Washington, DC, NTIS, no. N85-11903, Oct. 1984.
15. J. P. Silvester, M. T. Genuardi, and P. H. Klingbiel, *NASA's Online Machine Aided Indexing System,* NASA Contractor Report 4518, National Aeronautics and Space Administration, Washington, DC, NTIS no. N93-26901, Sept. 1993.
16. G. M. Hodge, "Functional Technologies for Database Reengineering: Technologies to Select and Apply," in *Impacts of Changing Production Technologies,* 1994 NFAIS Report Series, no. 3, D. Kaser, ed. National Federation of Abstracting and Information Services, Philadelphia, 1995, pp. 18–20.

17. G. M. Hodge, *Automated Support to Indexing,* 1992 NFAIS Report Series, no. 3, National Federation of Abstracting and Information Services, Philadelphia, 1992.
18. M. M. K. Hlava, *Machine Aided Indexing at Access Innovations, Inc.,* DTIC'95: Annual Users Meeting and Training Conference handout, Access Innovations, Albuquerque, NM, 1995.
19. G. Lustig and G. Knorz, *AIR/PHYS Pilot Application Project: Pilot Application of Automatic Indexing and Improved Retrieval Methods Using the PHYS Data Base,* Fachinformationszentrum, Energie Physik Mathematik GmbH, Karlsruhe, Federal Republic of Germany, 1986, pp. 1–30.

Some additional references for further reading on the subject of MAI are as follows:

Cavazza, M. and P. Zweigenbaum, "Extracting Implicit Information from Free Text Technical Reports." *Inform. Proc. Mgt.,* **28**(5), 609–618 (1992).

Come, C., ed., "Knowledge Bases to Improve Access to Documents," in *La Lettre de l'INIST,* no. 7, Oct. 1995, ISSN 1250-5943.

Edwards, S., "Investigation of a Computer-Assisted Indexing System for Its Practical Application in a Production Environment," paper presented at the 56th Annual Meeting of the American Society for Information Science, Oct. 1993.

Fidel, R., T. Bellardo Hahn, E. M. Rasmussen, and P. J. Smith, eds., *Challenges in Indexing Electronic Text and Images,* American Society for Information Science monograph series, Learned Information, Inc., Medford, NJ, 1994.

Hersh, W. R. and D. Hickam, "Information Retrieval in Medicine: The SAPHIRE Experience." *JASIS,* **46**(10), 743–747 (Dec. 1995).

Hodge, G. M. and J. L. Milstead, *Automated Support to Indexing,* 2nd ed., NFAIS Report Series, National Federation of Abstracting and Information Services, Philadelphia, 1997.

Humphrey, S. M. and D-C. Chien, *The MedIndEx: Research on Interactive Knowledge-Based Indexing and Knowledge Management,* National Library of Medicine, Bethesda, MD, NTIS no. PB90-234964/ AS, 1990.

Korzeniowski, P., "End-users at NASA Are the Link Between Data and Technology." *Infoworld,* **16**(37) (Sept. 19, 1994).

Milstead, J. L., "Methodologies for Subject Analysis in Bibliographic Databases." *Inform. Proc. Mgt.,* **28**(3), 407–431 (1992).

Minecci, C. M. and G. M. Hodge, "Machine-aided Indexing Productivity and Organizational Implications." *Inform. Serv. Use,* **8**, 133–138 (1988).

Seloff, G. A., "Automated Access to the NASA-JSC Image Archives." *Libr. Trends,* **38**(4) (1990).

Silvester, J. P. and P. H. Klingbiel, "An Operational System for Subject Switching Between Controlled Vocabularies." *Inform. Proc. Mgt.,* **29**(1) 47–59 (1993).

Sperberg-McQueen, C. M., *Guidelines for the Encoding and Interchanging of Machine-Readable Texts (P2),* Listserv TEI-L@UICVM.

JUNE P. SILVESTER